CLUSTERING
PROBABILISTIC HIERARCHICAL CONSUMER VIDEOS BY FINDING STRUCTURE IN
Abstract. Accurate and insightful video clustering and summarization are key to enabling effective video retrieval and recommendation. However, existing methods face challenges in dealing with the heterogeneity of video content and the dynamic nature of user preferences. To address these challenges, we propose a novel probabilistic hierarchical clustering method that incorporates user preferences and content similarity. The method aims to discover a hierarchy of video clusters that not only reflects the inherent structure of the video data but also captures the evolving preferences of users.

SUBMITTED FOR PUBLICATION

May 24, 2002

David Gauda-Perez
Alexander Liu
Ming-Ting Sun

Probabilistic Hierarchical Clustering
Finding Structure in Consumer Videos by
In this paper, we present a new framework for understanding the emotional expression of video content. In Section 1, we provide an overview of related work and introduce the concept of emotional expression in video content. Section 2 discusses the main objectives of this study. In Section 3, we describe the methodology used to analyze the emotional expression of video content. The models are based on a hierarchal, multi-modal approach, which includes models for (a) video analysis and (b) emotional expression analysis. In Section 4, we present the results of our experiments, which show that our approach is effective in capturing the emotional expression of video content.
Previous Work

People do not instantly recall the same scene twice. In other words, none video content tends to be

memorized in time.

In terms of dataset duration and number of shots per channel.

- People can focus their attention when watching only for a limited amount of time, which translates

special features:

- The structure of home videos is “self-similarity” to the structure of consumer still pictures.

- The structure of home videos is “self-similarity” to the structure of consumer still pictures.

- Some scenes are detected (e.g., zoom-in/out) by

- Some scenes are detected (e.g., zoom-in/out) by

- Some scenes are detected (e.g., zoom-in/out) by

- Some scenes are detected (e.g., zoom-in/out) by

- Some scenes are detected (e.g., zoom-in/out) by

- Some scenes are detected (e.g., zoom-in/out) by

Several characteristics distinguish home videos from other video sources.

What is Consumer Video?

In some consumer scenes, performance evaluation and the results of our methodologies are discussed in Section 8. Finally, Section 9 draws some concluding remarks.

Performance evaluation and the results of our methodologies are discussed in Section 8. Finally, Section 9 draws some concluding remarks.
Analyzing the Cluster Structure of Consumer Video

1. The Rokak: Consumer Video Database

4. Analyzing the Cluster Structure of Consumer Video

1. 4.1. The Effect of Lumped Focus-Attention

2. 4.2. Analyzing the Cluster Structure in Home Videos

An interesting question is how to classify the different types of viewers based on their viewing patterns. This can be achieved by identifying clusters of viewers with similar viewing behaviors. By doing so, we can gain insights into the preferences and interests of different groups of viewers. This information can be used to tailor marketing strategies and content recommendations to specific segments of the audience.

Furthermore, by analyzing the viewing patterns of different clusters, we can understand how they differ in their preferences and interests. This can help us identify new markets and opportunities for content providers. Additionally, it can assist in understanding the impact of different factors on viewer engagement and retention, which can inform strategies for improving user experience and increasing viewer satisfaction.

In conclusion, analyzing the cluster structure of consumer video data provides valuable insights into viewer behavior and preferences. This information can be utilized to make data-driven decisions in the realm of content creation, marketing, and user experience design.
The complete information is the distribution of the number of stops per video duration, denoted by $p(\tau)$.

**Figure 1**: Empirical distribution of normalized shot duration, and the GMM approximation. The maximum cluster duration is 1217 s.

As a consequence of lack of attention, very long video clusters are rare.

Where $N$ is the number of components in the mixture, $\alpha$ denotes the prior probability of the $i$-th component, and $p(\tau)$ is the empirical distribution of the duration of shots.

$$
(1) \quad \sum_{i=1}^{N} \alpha_i p(\tau | \theta_i) = \sum_{i=1}^{N} \alpha_i \int p(\tau | \theta_i) d\tau
$$

In empirical distribution and an approximation by a GMM, we have made use of this information. Let $e \in \mathcal{E}$ denote short duration. $P(z)$ is the number of clusters and $N$ is the number of shots. The $i$-th component of the mixture has a duration of short shots. $\alpha_i$ is a combination of short video shots added to the mixture. If we can see that the clusters have a duration of the empirical distribution of the number of clusters.

We assume that not only shots but also shots with clusters have a duration of clusters. However, this feature was not used in the analysis.

The GMM mixture is deeply related to the creation of mixture approximations.
of the clusters are composed of six or less shots.

Figure 2. Empirical distributions of (a) number of shots per sequence (b) number of clusters per sequence.

Number of Shots

Frequency

Number of Shots per Video Sequence

Number of Clusters

Probability Density Function

Distributions of Shots per Cluster

Number of Shots per Cluster

Probability Density Function

Number of Clusters per Video Sequence
4.2.2 The Effect of Continuity

While the benefits of computational similarity can be applied for characterization, clusters are grouped in time and the clusters should be characteristic of those clusters. In our problem, a rotation in the trend is one of the visual features of interest. In our experiment, only about 30% of the clusters possess this characteristic. In other words, home video clusters composed of non-referential shots (dominant and background elements) are preferred.
Video Segmentation Extraction and Selection

The selection of models for the distribution of the results are described in the next section. Our methodology is summarized in Fig. 1. The automatic determination of an optimal feature space and

\[
\text{argmax}_{x, y} f(x, y)
\]

where \( f(x, y) \) is the class conditional probability of the observed features, given \( x \) and the

\[
P(y|x) = \frac{P(x|y)P(y)}{P(x)}
\]

...the class that must be selected

Video Segments

2 Our Approach: Probabilistic Hierarchical Clustering of Home
These are simple estimations of multivariate distributions. In this paper, we have selected joint histograms \( f(x,y) \) for representations that have been proposed for content-based retrieval. In the second phase, from the various images and scenes, the set of shots is retrieved. In this phase, clusters are sequentially detected using a technique similar to that described,

6.1 Extraction of Visual Features

![Diagram of video extraction process]
6.2 Section of Visual Features

By [2] the space was reduced, and a measure of similarity could be defined if the dimensionality of the

Alternative, principled measures of similarity can be defined if the dimensionality of the

where the index T indicates the rank for (I).

(2)

'\{ Y_{n}^* \} \in \{ T_{n}\}, 1 \leq Y_{n}^* \in A_{n}^{(f_{S} \times f_{Y} \times q)} p_{Y_{n}^*} = (f_{S} \times f_{Y} \times q) p_{Y_{n}^*}

where \( A_{n}^{(f_{S} \times f_{Y} \times q)} \) is a set of the number of the Yt segments in the image. The measure of similarity between two images is defined as

\( \text{similarity} = \frac{\sum_{i=1}^{N} \left( Y_{n}^* \right)}{N} \)

\( \text{similarity} \) is a measure of the similarity between two images based on the correlation coefficient. The measure is defined as

\( \text{similarity} = \frac{\sum_{i=1}^{N} \left( Y_{n}^* \right)}{N} \)

Other features can be reduced and added in a straightforward fashion.

2. Edge-based features, including edge density and edge directions

3. Color features (known to be illumination-invariant), non-intensity quantized to 32 levels

The combined color and scene structure information, and these show to significantly improve color-only

We estimated the distributions of the two-class conditional

\( P(Y_{n}^* | X_{n}) = \sum_{i=1}^{N} \left( Y_{n}^* \right) \)
based on the Bhattacharyya coefficient.

Figure 6: Visual similarity features with subject detection and random frame extraction.
The joint class-conditioned pdf of the observed features are represented by multivariate GMMs.

The data analysis: Modeling of Likelihood Functions and Prior

The accuracy of likelihood functions will become less likely to drop to the same value as a result of the

| Accurate Selection of Temporal Features

Table 1: Feature Selection T' metric

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Feature</th>
<th>T' Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.29X</td>
<td>RCB-REDEN</td>
<td>0.780</td>
</tr>
<tr>
<td>0.29X</td>
<td>RCB-REDEN</td>
<td>0.730</td>
</tr>
<tr>
<td>0.29X</td>
<td>RCB-REDEN</td>
<td>0.690</td>
</tr>
</tbody>
</table>

Table 2: Comparison of Similarity Measures: John Hesgen REDDEN

<table>
<thead>
<tr>
<th>Measure</th>
<th>RCB-REDEN</th>
<th>RCB-REDEN</th>
<th>RCB-REDEN</th>
<th>RCB-REDEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.29X</td>
<td>0.780</td>
<td>0.730</td>
<td>0.690</td>
<td>0.650</td>
</tr>
<tr>
<td>0.29X</td>
<td>0.790</td>
<td>0.740</td>
<td>0.700</td>
<td>0.660</td>
</tr>
<tr>
<td>0.29X</td>
<td>0.800</td>
<td>0.750</td>
<td>0.710</td>
<td>0.670</td>
</tr>
</tbody>
</table>
8 Experiments and Results

where \( f(x) = 1 \) if the \( x \)-th training sample belongs to the class \( y \) and zero otherwise.

\[
\min_{\theta, \zeta} 
\sum_{i=1}^{N} \left[ \frac{1}{2} d^{2}(y_{i}, \xi \theta) + \frac{1}{2} \left( 1 - \xi \right) \nabla_{\zeta}^{T} \nabla_{\zeta} \right] = \mathcal{L}(\xi, \theta)
\]
8.2 Performance Evaluation Procedure

To assess the effectiveness of the proposed algorithm, we conduct a series of experiments in a controlled environment, and then compare the results to those obtained using the proposed algorithm. Our experiments include two main components: (1) evaluating the accuracy of the proposed algorithm, and (2) comparing the performance of the proposed algorithm with existing methods.

Table 1: Performance Evaluation Procedure of Semantic Document Retrieval

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Evaluation</th>
<th>Generation</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>7</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>4</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>14</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>16</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>12</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>6</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>8</td>
</tr>
</tbody>
</table>

8.3 Ground-truth

The results are compared with the proposed method and the baseline method. The ground-truth is obtained by human evaluation of a set of relevant documents.

Table 2: Performance Evaluation Procedure of Semantic Document Retrieval

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Evaluation</th>
<th>Generation</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>7</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>4</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>14</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>16</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>12</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>6</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>8</td>
</tr>
</tbody>
</table>
8.3 Results

Independent variable: Results are then accumulated and averaged over the whole database.

where \( N \) stands for the number of dops, \( A \) for the \( \psi \), and \( T \) for the gap.

\[\frac{\Delta N}{\Delta T} = c \Delta \psi + \Delta T \]

and \( \alpha = \frac{\Delta N}{\Delta A} \), the expressions are

\[\frac{\Delta N}{\Delta A} = \alpha \psi \]

Additionally, we present both figures for comparison. For macro-averages, \( \Delta N = 0 \), the figures are
We use the Bayesian approach to specify a prior on the probability of such in error, and include a method of assessing the probability. The true number of correct answers is observed for the particular of false negatives (to). The mean number of correct answers is determined to be a function of the true number of correct answers. A similar result can be obtained for the true number of correct answers.

The results show that our method outperformed all of the baseline methods, even with the additional process of inferring the correct number of clusters. Using macro-averages (F-measure, classification accuracy), the method performed better in most cases.
The effect of the prior distribution is shown in Table 2. The use of a uniform prior does not affect the posterior distribution of the probability of correcting operations.

### Table 2: Shot Assignment Performance

|  | 0.025 | 0.075 | 0.125 | 0.175 | 0.225 | 0.275 | 0.325 | 0.375 | 0.425 | 0.475 | 0.525 | 0.575 | 0.625 | 0.675 | 0.725 | 0.775 | 0.825 | 0.875 | 0.925 | 0.975 |
| Method | 0.2300 | 0.2400 | 0.2500 | 0.2600 | 0.2700 | 0.2800 | 0.2900 | 0.3000 | 0.3100 | 0.3200 | 0.3300 | 0.3400 | 0.3500 | 0.3600 | 0.3700 | 0.3800 | 0.3900 | 0.4000 | 0.4100 |
| Posterior CDF | 0.9188 | 0.9184 | 0.9180 | 0.9176 | 0.9172 | 0.9168 | 0.9164 | 0.9160 | 0.9156 | 0.9152 | 0.9148 | 0.9144 | 0.9140 | 0.9136 | 0.9132 | 0.9128 | 0.9124 | 0.9120 | 0.9116 |
| Prior CDF | 0.9190 | 0.9185 | 0.9180 | 0.9175 | 0.9170 | 0.9165 | 0.9160 | 0.9155 | 0.9150 | 0.9145 | 0.9140 | 0.9135 | 0.9130 | 0.9125 | 0.9120 | 0.9115 | 0.9110 | 0.9105 | 0.9100 | 0.9095 |
| Posterior CDF | 0.9192 | 0.9187 | 0.9182 | 0.9177 | 0.9172 | 0.9167 | 0.9162 | 0.9157 | 0.9152 | 0.9147 | 0.9142 | 0.9137 | 0.9132 | 0.9127 | 0.9122 | 0.9117 | 0.9112 | 0.9107 | 0.9102 | 0.9097 |
| Prior CDF | 0.9194 | 0.9189 | 0.9184 | 0.9179 | 0.9174 | 0.9169 | 0.9164 | 0.9159 | 0.9154 | 0.9149 | 0.9144 | 0.9139 | 0.9134 | 0.9129 | 0.9124 | 0.9119 | 0.9114 | 0.9109 | 0.9104 | 0.9099 |

### Bayes’ Rule

\[
\text{Posterior CDF} = \frac{\text{Likelihood} \times \text{Prior CDF}}{\text{Evidence}}.
\]
9 Conclusion

To extract these features from the video streams, we are considering the application of two different methodologies: one based on the estimation of motion patterns and another based on the detection of visual saliency. These methodologies have been developed in order to detect the most informative regions of the video content.

8.4 Limitations

On the video streams for each burst, we are considering the application of two different methodologies: one based on the estimation of motion patterns and another based on the detection of visual saliency. These methodologies have been developed in order to detect the most informative regions of the video content.

7.2 Effect of Audio Probability

<table>
<thead>
<tr>
<th>Audio Probability</th>
<th>0.036</th>
<th>0.060</th>
<th>0.090</th>
<th>0.120</th>
<th>0.150</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected prior</td>
<td>0.384</td>
<td>0.320</td>
<td>0.256</td>
<td>0.192</td>
<td>0.128</td>
</tr>
</tbody>
</table>

The evaluation of the proposed methodology was performed using a dataset of video streams, and the results indicate that the proposed methodology is effective in extracting the most informative regions of the video content.
Figure 10: General structure on a board video sequence (detail). Each shot is represented by one frame.

Figure 15: General structure on a frame video sequence (detail). Each shot is represented by one frame.
Methods: (a-d) Frames extracted from parts of video clips, that were categorized manually by our

Figure 12: (a-d) Frames extracted from parts of video clips, that were categorized manually by our

Figure 11: Displaying the video segments trees.

Figure 11: Displaying the video segments trees.
Acknowledgements

(1)

\[(\text{corr} + \text{corr} + \text{corr} + \text{corr}) = \text{corr} \]
\[(\text{corr} + \text{corr}) = \text{corr} \]
\[(\text{corr} + \text{corr}) = \text{corr} \]
\[(\text{corr} + \text{corr}) = \text{corr} \]

\[\{T \cdots 1 \in I \cap I \cap \cdots \} \cap \{Y \cdots 1 \in \cdots \} = 1\]

\[\{T + Y \cdots 1 \in \cdots \} \cap \cdots \} = 1\]

When a part of sentences is inserted, the model of the new sentence is updated by

(b) Replace the MIP criterion to merge the part of sentences (Eq. 5).

(c) If the sentences are merged (i.e., the model of the merged sentence, then

Apply the MIP criterion to merge the part of sentences (Eq. 5).

(d) Extract all sentences from the chosen. This part of sentences is the one that has the highest

2. Queue the MIP criterion. Until the queue is empty.

I. Queue initialization. At the beginning of the process, all sentences are composed of one short

The method consists of two steps: queue initialization and queue updating/merging.

Appendix 10


